DOCUMENT RESUME

ED 233 996 SP 022 900

AUTHOR Peterson, Ken; Kauchak, Don
TITLE Teacher Evaluation: Perspectives, Practices, and Promises.
INSTITUTION Utah Univ., Salt Lake City. Center for Educational Practice.
PUB DATE Jan 82
NOTE 53p.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Formative Evaluation; *Legal Problems; Observation; *Public School Teachers; Self Evaluation (Individuals); Student Evaluation of Teacher Performance; Summative Evaluation; Teacher Characteristics; *Teacher Evaluation

ABSTRACT

This report highlights major issues, techniques, and directions in the evaluation of public school teachers. The paper begins by setting a perspective on the process of, and needs for, evaluation. The main body of the report is devoted to a summary and critique of various teacher evaluation methods. A discussion is given of the efficacy of, and problems involved in, evaluating with certain techniques: (1) credentials; (2) personal characteristics; (3) student outcomes (pupil achievement); (4) classroom visits; (5) self-reports; (6) student reports; (7) peer review; (8) competency-based teacher evaluation; and (9) systematic observation. Legal issues involved in teacher evaluation are discussed in the third section. The final section of the report suggests a variety of approaches and strategies that may be combined to result in more satisfactory teacher evaluation; a bibliography is appended. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED233996

Center for Educational Practice

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Don Kauchak
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
✓ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

SP 022 900

Graduate School of Education / University of Utah

FORWARD

The purpose of this report is to highlight major issues, techniques, and directions in the evaluation of public school teachers. The paper begins by setting a perspective on the process of, and needs for, evaluation. Next is the main part of the report: a summary and critique of various teacher evaluation methods. Some of these techniques are difficult to defend in terms of objectivity and fairness, but others provide a great deal of useful information if implemented carefully. The third section of this paper presents the most important legal considerations in evaluating teachers. Finally, suggestions for future development in teacher evaluation are discussed.

This paper has a number of necessary limitations. First, it has a primary objective of stimulating discussion and giving initial direction to a wide variety of readers: legislators, public school teachers, university educators, and lay public groups interested in education. Therefore, the discussion will not have the degree of specialization which might be desired by any single group of readers. Second, the paper is too brief a survey to be comprehensive. However, citations and resource bibliography are included which will assist those who desire additional information on the evaluation of teachers.

TABLE OF CONTENTS

Forward 1

Introduction 3

Section I--Perspectives in the Evaluation of Teachers 6
A. Purpose of Evaluation 7
B. Formative and Summative Evaluation 7
C. Quantitative and Qualitative Evaluation Data 8
D. Multiple Data Sources 8
E. Audiences for Teacher Evaluation 9

Section II--Techniques for the Evaluation of Teachers 11
A. Credentials 12
B. Personal Characteristics 13
C. Student Outcome (Pupil Achievement) 14
D. Classroom Visits 18
E. Self-report 23
F. Student Reports 25
G. Peer Review 27
H. Competency-Based Teacher Evaluation 29
I. Systematic Observation 34

Section III--Legal Issues in Teacher Evaluation 39

Section IV--Directions for Development of Teacher Evaluation 43

Bibliography 46

INTRODUCTION

The evaluation of public school teachers is of major significance for legislators, school administrators, students, parents, teacher-preparation universities, the lay public, and for teachers themselves. Teacher evaluation has implications for quality, accountability, training, and the well-being of teachers. Yet it is one of the most underdeveloped and ignored areas of educational development and research.

While many indirect sources of evidence suggest that public school teachers provide effective and efficient service, current teacher evaluation practices and structures do not provide satisfactory information for classroom teachers, lay public, legislators, or teacher-preparation institutions. At present, teacher evaluation consists predominantly of school principal rating of teacher performance and professional characteristics. Administrative evaluation is an important educational practice; however, it does not give authoritative direction to practitioners, verify to the public the quality of teaching in the classrooms, or provide specific information to universities for the improvement of teacher preparation. At present, teacher evaluation is characterized by tradition, uncertainty, confusion of control, and conflict among interested audiences.

Three major obstacles have prevented the development of effective teacher evaluation practices. The first problem is the state of the art: few practices and procedures presently exist which provide useful data

about the assessment of teachers and teaching. The second obstacle concerns the large number of audiences involved. Teacher evaluation data are needed by professionals, legislators, school administrators, the lay public, and teacher training institutions. These groups have not, in the past, worked together nor have they resolved their competing claims for information. A third problem of teacher evaluation development is that the vast majority of present work in the area has been preempted by administrative evaluation; other purposes have been neglected.

The scope of the need for valid teacher evaluation information and the obstacles to development preclude short-term solutions. A long-term research and development effort characterized by careful planning and involvement of the multiple audiences is needed. Specifically, successful development of teacher evaluation practices and structures should address the needs of at least the following groups:

- teachers, through their professional organization,
- local school districts,
- state legislators,
- state boards of education, and
- teacher education institutions.

Now is the crucial time to begin work on new practices and structures for teacher evaluation for several reasons. First, pressures have never been greater on teachers to give public evidence of the results of their efforts. Second, scarce public funds are sought with increasing competitiveness by a great number of public institutions. At the same time, current developments in curriculum evaluation, performance evaluation, and the sociology of professions provide promising practices which have yet to be tested in teacher evaluation. In short, increasing pressure and concurrent development of practices in other areas of education make it more likely that successful collaboration among the interested groups

is possible.

The development of comprehensive teacher evaluation systems requires an understanding of the evaluation process, its products, and the different evaluation formats. Some evaluation systems provide useful information for the improvement of instruction while others provide information about the question of quality at the school, district, and state level. A description of these various forms of evaluation are found in Section I, which presents background information about evaluation processes. Some readers may wish to skip Section I and proceed directly to analysis of methods (Section II), legal issues (Section III), or suggestions for development of teacher evaluation (Section IV).

SECTION I

PERSPECTIVES IN THE EVALUATION OF TEACHERS

THE PURPOSE OF EVALUATION

Evaluation is an activity which determines the worth, merit, or value of a performance, product, or person in a particular role. In distinction, research has the aim of determining ultimate, generalizable truth while teacher evaluation has the function of providing information for decision making for specific groups of people. The value of research findings lies in how closely they match reality; the value of evaluation results is determined by how adequate and satisfactory they are for the concrete deliberations of an actual audience. Specifically, the purpose of teacher evaluation is not to determine the question of what makes an ideal teacher (a question for research), but how good a given performance, product, or person has been in an actual situation. Most often this judgment is developed and considered in terms of comparison with other performances, products, or persons. There typically is not a single best method to evaluate teachers, but rather, some ways which are more satisfactory, adequate, and defensible for a given group and situation. This idea will be further developed in later sections.

FORMATIVE AND SUMMATIVE EVALUATION

Evaluation has two major uses. In formative evaluation, data are used as feedback for change and improvement. The second kind of evaluation, termed summative, results in employment decisions and similar judgments. This distinction is important because often very different techniques are used according to the intended purpose. In practice, it may be difficult or impossible to accomplish both types of evaluation at the same time. For example, a teacher will participate in one manner with the goal of improvement, as in formative evaluation, but will behave quite differently if a job is at stake, as in one kind of summative evaluation. The audience for the evaluation may have either or both uses in mind but should be clear

about what their intentions are. Summative evaluation techniques, because of their consequences, are usually more narrow in scope and thus require more rigor and systematic application in practice.

QUANTITATIVE AND QUALITATIVE EVALUATION DATA

Information used for evaluative decisions or judgments may appear in the form of numbers, such as test scores and ratings, or in the form of verbal descriptions, such as reports and comparisons. Most evaluations involve both kinds of data. The event which is being assessed determines the form of the data. For example, observers can rate some aspects of a teacher's performance on a numerical scale which can be compared with similar performances of other teachers. Other kinds of teacher performance, for example, the strategy of beginning a class, cannot easily be reduced to numbers; for these aspects we must rely on verbal descriptions. Most importantly, the kind of data gathered in evaluation must be appropriate and the best available. Judgments about the quality of data used in evaluation are based on validity, reliability, cost-effectiveness, absence of unwanted side-effects, long-term significance, and justice or fairness to participants (Scriven, 1973).

Problems arise with evaluation systems in which there is over-quantification--when decisions or judgments predominantly rely on satisfying numerical requirements. It is rare, in something as complicated as teaching, that a numerical decision by itself is adequate for judgment.

MULTIPLE DATA SOURCES

In making evaluation decisions, it is important to use as many different sources of data as possible. For example, the value of a new commercial product may be judged by sales records, cost-effectiveness in manufacturing and absence of undesireable side effects. The process of evaluating teaching

is an even more complicated task which requires correspondingly more data sources and indicators of quality. One complication of finding data sources is that the teacher is not entirely responsible for even the immediate outcome of her talents or efforts, e.g., student effort and prior achievement greatly determine the amount learned. In addition, only some goals of teaching are visible in the short term and/or are easily measured. Finally, much of what a teacher does is context-dependent; what works in one place with one kind of student is not good practice in another setting with another kind of learner.

Thus, it is important in evaluating teachers to use a variety of assessments of teacher quality and to balance and weigh these factors according to the goals of the evaluation. It is necessary to take care in gathering data so that the result is not merely the sum of parts; quality teaching exists in different patterns and is therefore evidenced through a variety of means. The second major section of this paper reviews a range of possible techniques for gathering evidence about teacher quality.

AUDIENCES FOR TEACHER EVALUATION

Teacher evaluation has a number of audiences with an interest in the resulting information and judgments: school administrators, teachers themselves, parents, voters, legislators, and teacher-training institutions. These audiences differ in their evaluative questions, the kinds of evidence which satisfy them, and even the language of evaluation. With some groups e.g., school administrators, needs and procedures are well known. For other audiences only scant information about specific questions and needs exists. For these groups development of effective teacher evaluation practices will require a more precise understanding of their needs for data and interpretation.

Audiences differ in evaluation needs because they have different roles and uses for teacher evaluation data. For example, state legislators have as a primary function the judicious use of taxpayers' money. In this respect, the teacher evaluation data of most interest to them help to answer questions such as "Is the taxpayer getting a good return for her education tax dollar?" and "Are there more cost-effective ways of allocating educational funds?" School administrators, by contrast, are more concerned with teacher evaluation data that provide information about the quality of their programs and teachers. In this sense, they are interested in both formative and summative teacher evaluation data which will be used to shape future decisions and to make personnel decisions. A third audience or consumer of teacher evaluation data is teachers themselves. As professionals, teachers need to know when their actions are effective and ways in which their teaching can be made more effective. In addition, there is increasing evidence that teacher satisfaction with the profession can be strengthened by availability of reassuring and respected feedback about effectiveness.

Another way of differentiating the evaluation needs of different audiences is in terms of scope. State legislators typically are not interested in teacher evaluation data dealing with individuals; instead they find broad, descriptive data comparing programs, districts, or states to be most helpful. By contrast, the most valuable data for teachers interested in improving their own teaching effectiveness must be quite specific and individualized.

These different perspectives on teacher evaluation suggest the need for using a variety of teacher evaluation practices and should be kept in mind as various teacher evaluation practices are discussed in the next section.

SECTION II

TECHNIQUES FOR THE EVALUATION OF TEACHERS

CREDENTIALS

Credentials are documentation of professional training, certificates, degrees, preservice or inservice credits, professional memberships, grade-point averages, and teaching experience. Since the completion of programs and the results of experience are presumed to result in more effective practice, the evaluative assumption often held is that the more credentials a person has, the better teacher he is.

In practice, credentials do not assist in evaluating the immediate, manifest quality of teachers. Associations between credentials and student learning have been found to be weak (Guthrie, 1970) or nonexistent (Rosenbloom , 1966) for several reasons. First, training programs and courses may be directed toward specific abilities which are not assessed by the measures of teaching effectiveness being used. Second, credentialed backgrounds do not affect in a systematic way any specific categories of behavior across populations of teachers. A third reason for the lack of a direct connection between credentials and quality is the individual nature of the teaching act. Each teacher operates in the classroom based on unique and perhaps idiosyncratic structures of teaching knowledge, skill, and attitude. It is extremely difficult to measure out individual, specific contributors to this underlying structure and to then demonstrate a relationship with concrete, manifest teaching performances. Experience, by itself, or a particular academic background, by itself, have not been found to be detectable contributors to teaching ability. This conclusion, in terms of teacher education, is not surprising; parallels exist in the law and medical professions. Although degrees in these areas assure minimal levels of competence, the level of quality within these licensed populations varies considerably.

While credentials are an ineffective evaluative measure <u>within</u> a group of teachers, credentials are not altogether irrelevant to the question of

teaching competency. The position that there is no relationship whatsoever between professional training (credentials) and quality of teaching is not defensible. According to this perspective, professional training makes no difference at all. This is not the case. Anyone who doubts that prepared people in fact do perform better in the classroom could carry out a study with large and diverse numbers of persons randomly assigned for a year to typical classrooms. Lay teachers have been systematically tested only in very unrealistic teaching situations with limited time and objectives, small populations of students, and pre-selected materials (Popham, 1971).

Another consideration is the practice of paying teachers for their experience or for completion of degrees and units of inservice credit. This is defensible not in terms of specific outcomes but in the recognition that professional development is presumed to contribute to the underlying structures of teachers. In the absence of more effective teacher evaluation practices, problems with direct relationship are not at this time sufficient argument to disband this practice.

PERSONAL CHARACTERISTICS

When teachers are evaluated in terms of personal characteristics, items such as intelligence, prior experience, friendliness, tact, style, language, humor, energy, stability, caring, grooming, dress, punctuality, and patience are considered. These characteristics have great appeal as evaluative criteria because most people have strong, clear opinions about them and assume them to be easily recognizable in an individual. Presumably, the characteristics of a person ought to greatly affect learning, the classroom atmosphere, and the general effectiveness of a professional.

The problems involved in using personal characteristics in teacher evaluation are twofold. One is determining which personal characteristics

are important and productive; the second is to reach agreement on how they should be measured. Empirical data do not support the idea that personal characteristics are in fact linked to pupil performance (McNeil and Popham, 1973).

The second problem, objectively measuring the existence of these personal characteristics, also poses difficulties. How can one tell if a given teacher is "dynamic" or has a "sense of humor?" There is little agreement among people in judging characteristics: the traits and their effects are "in the eye of the beholder" (McNeil & Popham, 1973). Characteristics which appear obvious to one observer are interpreted quite differently by others. In controlled studies of such judgments, rater agreement is actually very low (Cook & Richards, 1972). This is especially the case when the perceptions of students and adults are compared (Peterson & Yaakobi, 1980).

Despite these problems, Ingils (1970) reported that the use of personal characteristics in evaluating teachers is a common strategy. It often is the case that administrators are required to maintain appearances at schools and consequently feel justification for using rating systems that include personal characteristics. Even though directly observable personal characteristics, such as acceptable dress standards, can be useful in guaranteeing minimal adherence to district policies, it should be emphasized that hundreds of studies have failed to demonstrate relationships among teachers' knowledge, their personal characteristics, and their teaching effectiveness (Schalock, 1981). This fact strongly argues against use of personal characteristics for summative purposes, except in the most extreme cases, e.g., continued non-adherence to district policies.

STUDENT OUTCOME (PUPIL ACHIEVEMENT)

The amount students learn from a teacher is an evaluation criterion

which has great initial appeal to many who have considered the problem of teacher assessment. Many contend that the purpose of teaching is to produce the greatest achievement gains in students and that the value of a teacher is demonstrated by the learning of his students. Following this line of reasoning, some authorities have advocated reliance on student achievement as a prime determinant of teacher quality (e.g., Kerlinger, 1971). Closer examination of this approach reveals severe problems.

Three major obstacles to using student achievement in teacher evaluation have led to what experts call the "disasterous," "egregious," and "indefensible" use of achievement data for evaluating teachers (e.g., Glass, 1974). The first set of problems surrounds the logical connections between teacher performance and student outcome. The second area pertains to technical difficulties in the measurement of student gains. The third obstacle is the effect outcome systems have on educational programs. Student performance evaluation systems definitely affect the way a teacher acts in the classroom not always for the benefit of students.

Teacher quality and efforts are not always directly tied to student learning. For example, lack of student effort can thwart the effects of the most brilliant teachers. In addition, research has shown that parental expectations, prior achievement, socioeconomic status, and the general intellectual quality of the home all may have greater influence on pupil learning than does the teacher (Borich, 1977). Many school factors which are beyond the control of teachers have also been shown to affect pupil growth. These include classroom resources, number of students, and learning environments, such as the size of the room. Finally, teacher effects vary in their potency according to the age of students and the nature of the material which is to be learned; it is not fair to compare teachers who have different teaching assignments.

The technical problems of accurately measuring student learning for the purposes of evaluating teachers seem insurmountable at present. The five major problems here are:

1. What is to be tested is not clear.
2. Good (useable, valid, reliable) tests for summative evaluation purposes are not widely available.
3. Administration of these tests, for summative evaluation purposes, is difficult and expensive.
4. Gain data, not merely end-of-instruction achievement, are needed--and hard to get.
5. Stability of teacher influences is low.

These problems will be discussed in the paragraphs that follow.

There is general agreement on what teachers should be doing: students should be learning subject matter which consists of information, skills, and attitudes. At the same time, they should recognize their increasing competence, feel better about themselves, become better citizens, develop more responsibility, increase in problem-solving ability, prepare for a world of work, and develop independence. When these additional important goals of education are considered, it becomes difficult to narrowly specify and measure the job of a teacher. Fairness demands that a teacher be evaluated on the basis of the total job expectations rather than just a narrow segment of it.

Even if the purposes of teaching were fairly narrow and agreed upon, there still is the problem of a lack of good achievement tests for all levels and topics. The tests which do exist are very useful for pupil diagnosis, feedback for learning, promotion, and qualification for further study; they are not good for the purpose of absolute statements of pupil learning. A large part of the problem is the discrepancy between the content of standardized tests constructed at the national level and the goals of individual schools or districts. Typically, the tests measure

outcomes that are different from the goals of the teachers and do not measure what teachers were assigned to teach. Locally constructed achievement tests offer one solution to this problem but are difficult and expensive to construct and are not generalizable to other settings.

Even where valid achievement tests are available, they need to be well administered if summative decisions are to be made. In practice, large scale testing requires expert administrators and well controlled testing conditions. These are expensive and, in reality, are difficult to insure.

An additional problem in the use of achievement test scores to evaluate teachers is the selection of the test scores for analysis. The worst practice, at present, with achievement tests is the use of post-test only scores. The pertinent measurement is the amount a student learns from a class. However, a post-test score is influenced by prior achievement levels of the students, their individual abilities, and the resources available to the teacher during the class. It is patently unfair to compare or judge teachers without estimating the percentage of final achievement resulting from factors outside the control of the teacher. Even if gain data are sought, they are difficult to determine with reliability. As Borich (1977) has pointed out, if both the pretest and posttest have reliabilities of .80*, and the correlation between pretest and gain is .70 (both of these coefficients are common and expected values in education), then the resulting reliability of the gain score will be .33. Even more elaborate statistical techniques (residualized gain scores) will rarely approach necessary lower limits of reliability. While this practice may be defensible for research studies, it has never been done for purposes of general teacher evaluation.

---

*Rules of thumb for reliability coefficients: above .92 if individual educational decisions are to be made about students, above .80 if group decisions are to be made (e.g., curriculum) and down to .70 for research purposes if alternatives are not available.

The final measurement problem has to do with the stability of teacher effects. What a teacher does in one instance is not necessarily what he will do in another situation. Estimates of reliability for teacher effects range from .08 to .30 across two educational settings (Rosenshine, 1970). In order to generalize about a teacher's performance, teachers need to be observed in at least five situations with more than fifteen students in each situation. This would be impossible to accomplish within a year for elementary teachers who have only one class.

Even if the above measurement problems are dealt with, there remains a third major obstacle to the use of achievement test scores as the sole determinant of teacher effectiveness. This obstacle is the narrowing of focus in classrooms where these systems are in use. Teachers begin to teach to the test and to emphasize a specific expression of learning to the detriment of the broad scope and goals of most school subjects. Often ignored are difficult to measure educational goals such as personal initiative, aesthetic growth, and problem-solving ability. Reliance on achievement tests may tend to make teaching and learning trivial and rigid. Clearly, great care is required to use achievement tests without threatening the total educational program.

Taking all of the problems of achievement test results into account, they do not present much promise as a major criterion in teacher evaluation. For this reason, the National Educational Association has publicly disavowed any evaluation system which employs them. While this position might be overstated, it does illustrate the polarity of views on this controversial subject.

CLASSROOM VISITS

Evaluation through classroom visits employs short-term, data-gathering visits by administrators, supervisors, or peers. The use of classroom visits

is based on the idea that the best way to evaluate the quality of a teacher is to see that person in action. Proponents of classroom visits point out that this practice provides an opportunity to assess the climate, rapport, interaction, and functioning of the classroom as no other data source can (Evertson & Holly, 1981). Due to ease of administration and a long history of use, classroom visits remain a mainstay of teacher evaluation practice (Ingils, 1970). It should be noted that classroom visits differ from systematic observations (see the section on this topic) in that they do not use trained and monitored observers, reliable sampling, limited and validated observation categories, and standard recording procedures.

Classroom visits can serve some important and needed administrative functions. For example, they indirectly insure classroom control, serve as a check on a good number of district guidelines for teachers, provide for the visitor to become more familiar with a teacher's work, and check on the appearances of classrooms for order and neatness.

In current practice, classroom visits are the main strategy for teacher evaluation. Their limited scope presents obvious problems in the evaluation and improvement of teacher performance (Evertson and Holly, 1981). Teachers who must rely on administrator visits as the main or only source of evaluative information are placed in a position where this power can interfere with the leadership functions of their principals. At the same time, such visits do not provide adequate and reliable data about teacher performance for many audiences, including teachers themselves, the lay public, and teacher preparation institutions. Finally, empirical studies have not found administrator ratings to be related to pupil learning (Medley & Mitzel, 1959).

A main problem with classroom visits is reliability. A reliable evaluation is one in which several persons agree about the same class, or one evaluator reports the same class results time after time. While most

people, professional educators and lay persons alike, intuitively feel that they can assess teacher quality just by watching for a while, empirical tests consistently show how little agreement (i.e., reliability) is derived from classroom visits. Classrooms are very complicated places, and they change over time. For these reasons, many visits are required to observe them in a representative manner. The second reason for unreliability is within the observer herself. Due to biases, lack of perception of all that occurs, and a limited personal perspective, much that is relevant is missed and that which is noted falls within a personal frame of reference.

Cook & Richards reported a large scale classroom visit study in which principals and college supervisors rated 236 beginning teachers on a range of 23 personal and professional characteristics (e.g., "tact," "techniques of teaching"). A thorough analysis of the data revealed that "...the rating scales generated data that were more a reflection of the raters' point of view [role] than of a teacher's actual classroom behavior." (1972, p. 14).

Peer visits have been less studied than have administrator visits, but preliminary results suggest an equally poor performance. Centra (1975) studied college peer reviews in which two or three colleagues visited classrooms for two or three visits per quarter. Since this study took place at a new college, bias from political or friendship considerations was minimized. Centra found that the peer reports were generally unreliable. Correlation coefficients of interrater agreement were around .30. Correlations of individual items of observation ranged from zero to .45 at the highest. The high items were for visible factors, such as "uses examples during instruction." Other factors, such as "understood level of learning," were near zero.

The reasons for low reliability in classroom visits are complex but can be explained (Scriven, 1981). First, since the number of visits are

few, the apparent patterns are more likely to come from the observer than from the classroom itself. Second, the visitor focuses his observations according to the situation and his own personal interests; what he notices reflects his personal viewpoint. Third, because the recording system is inadequate, the observer relies on her recollections which are greatly determined by preexisting conceptions. Fourth, the relationship of the observer and teacher in terms of politics or friendships is important. It also is the case that preferences for personal style are too often emphasized. Finally, the act of visiting itself alters the teaching and the behavior of students in the classroom. Taken all together, these factors result in role-dominated reports.

Classroom visits by the building principal using a checklist is the most common evaluative technique (Ingils, 1970). It has appeal because of minimal expense, existing power relationships within the school, and apparent validity. It also has the tradition of the principal as instructional leader. Principal visits have a long history of use; they are legally strong because of precedent. The question of the soundness of classroom visits as an evaluation technique is overlooked because of their widespread use.

Classroom visits with checklists suffer from many problems. Most checklists in use combine characteristics (e.g., "enthusiasm"), difficult to observe inferences (e.g., "keeps interest up in students"), and items of inference and tradition which are not tied to student learning (e.g., "has everyone's attention before beginning"). In addition, rating forms often have both formative and summative uses, which interferes with the function of either intention. Many forms are overwhelming in their numbers of items for observation; some require response to 60 or more

topics. Recording procedures may be confusing, for example some items need a frequency count (e.g., "supportive statements") while others require a single check off notation (e.g., "used advanced organizers"). The vast majority of rating forms in current teacher evaluation use have not been checked for reliability. Data analysis of these forms often presents information as discrete categories and does not summarize findings so that readers can get a clear picture of what the observer saw. Finally, the conceptual foundation of most forms is lacking and, in part, invalid in terms of what is known about effective teaching.

An additional problem with classroom visits is their validity, i.e., demonstrated relationships between observed teacher actions and achievement gains or affective growth. Evertson and Holley, in a review of a number of studies on classroom visits, concluded that "...there is a fairly consistent failure to find relationship between ratings of teacher performance and other external measures of competence," (1981, p. 96). This conclusion is reinforced by Travers (1981) and Coker, Medley, and Soar (1980). The latter researchers investigated the relationship between a number of observable variables and student achievement and student self-concept. The study was conducted in 100 Georgia classrooms, ranging from first through twelfth grade. The researchers found that some teachers' behaviors were positively related to student achievement at some levels but not at others; similar findings were found for self-concept scores. In addition, certain teacher behaviors were positively related to student gains at certain levels and negatively related at others. These results suggest that observation systems cannot be designed for use with all teachers, at all levels, in all subjects. The link between teacher actions (as measured by classroom observations) and student achievement appears to be more complex than implied by the broad observation systems used

in classroom visit checklists. (See the section on Systematic Observation).

The above considerations suggest that classroom visits be limited to specific administrator needs and not take on the burden of the entirety of teacher evaluation (Evertson & Holly, 1981). Scriven (1981) proposes that visits can assess major deviations from teaching practice, such as accuracy of information, sexist or racist statements, immoral behavior, or complete lack of classroom discipline. However, the bulk of what is currently expected from classroom visits needs to be accomplished through the use of systematic observation, which is discussed in another section of this report.

SELF-REPORT

Self-assessment is an expected part of teachers' professional performance and can provide useful information. Though helpful for formative purposes, self-reports have great limitations for most types of summative teacher evaluation (McNeil & Popham, 1973; Carroll, 1981).

Research has provided a good deal of information about teacher self-reports. Teachers consistently monitor their own behavior in relation to goals, expectations, and outcomes (Festinger, 1954; Simpson, 1966) and are more likely to act on self-gained data than on information from other sources (Centra, 1972). Instructors have been shown to demonstrate significant improvement in subsequent student ratings when moderate discrepancies are identified between initial student ratings and instructor self-reports (Carroll, 1981). Finally, researchers have found that teachers can become more effective at self-assessment if training and opportunity to use self-reports were more available (Weiner & Kukla, 1970).

Self-reports can be valuable for several teacher evaluation purposes. Teachers, because of professional knowledge, can suggest categories of

performance and relations among teaching tasks and, in general, give a perspective on teaching performance which is informative to data collectors (Centra, 1977). Teacher self-assessment can also be of great value to administrators in helping to make teaching assignments which are satisfactory and productive.

The two major problems with wide use of self-report data in teacher evaluation are subjectivity which produces inaccurate data, not in agreement with objective data, and conflict of interest, especially for summative judgments (McNeil & Popham, 1973).

Empirical studies have generally demonstrated that self-ratings show little agreement with students' ratings. In a study involving 343 teachers from 5 colleges, Centra (1972) found a median correlation of .21 between self- and student ratings. In this study Centra also found a tendency for teachers to give themselves better ratings than did their students. Blackburn and Clark (1975) found little agreement between faculty self-ratings of teaching effectiveness and ratings by students, colleagues, and administrators. Significantly, these latter three groups did substantially agree on their ratings of the teachers. Peterson and Yaakobi (1980) reported a study of high school classrooms in which student reports and teacher self-descriptions of classroom behaviors had a mean correlation of .30. They also found that teachers' reports were inflated relative to student assessments. It may be the case that an optimistic view of one's self as a teacher, although unrealistic, is essential to performing the role.

Self-interest precludes the use of self-reports in most summative evaluation. Persons should not be expected to objectively contribute to final decisions about salary, retention, or promotion.

STUDENT REPORTS

Important, useful, and reliable data for teacher evaluation can be obtained through student reports of teachers. Student ratings produce a main source of information regarding the development of motivation in the classroom and the degree of rapport and communication developed between teacher and student. In addition, student ratings provide unobtrusive information on course elements such as textbooks, tests, and homework. Students are good sources of information about their instructors because they know their own case well, they have closely and recently observed a number of teachers, they maintain a unique position and perspective in comparison with other observers, and they benefit directly from good teaching.

Student reports are defensible sources of information about the performance of teachers for several reasons. The availability of a large number of students for use as data sources increases the reliability of their reports for many kinds of teacher observations. Reliabilities in the .8 to .9 and above range are quite frequent in the literature. Student report data, most often obtained through questionnaires, are relatively inexpensive to obtain in terms of time and personnel; data summarization is the major cost. In addition, student reports can be justified in terms of the viewpoint of students as consumers (McKeachie, 1979).

Student rating of instructors is one of the most heavily researched topics in teacher evaluation. The results of this inquiry are positive in their implications for teacher evaluation practice (Aleamoni, 1981; McNeil & Popham, 1973; Haak, Kleiber & Peck, 1972; Centra, 1980). Researchers found that student ratings of teachers are consistent among students and reliable from one year to the next. Studies also show that students can

successfully differentiate between teaching effectiveness and other affective dimensions such as attitude, interest, and friendliness of the teacher. Student ratings are neither capricious nor whimsical; students can consistently differentiate among instructors, and ratings are not based solely on popularity factors, a fear which is frequently expressed by teachers. Perhaps the most compelling argument for the use of student ratings is the fact that they do relate to the amount learned in a course. In a comprehensive analysis of forty-one studies reporting on 68 courses having multiple sections, Cohen (1981) found the mean correlation between the overall instructor rating and student achievement to be .43, the mean correlation between the overall course ratings and student achievement was .47. Significantly, Cohen found that these results were not affected by the type of institution or the type of class; these results were consistent in hard and soft disciplines, in pure and applied areas, and in life studies as well as other content areas. In addition, Aleamoni (1981) found that student ratings were positively related to colleague ratings, expert external judge ratings, and graduating seniors and alumni ratings.

In addition to summative purposes, student reports have been shown to be useful for formative evaluation functions. Tuckman and Oliver (1968) found instances in which supervisor ratings produced negative reactions in teachers while student reports of the same topics were positively received by teachers.

In the area of student evaluations, the bulk of research has been conducted at the college level. A number of studies, however, suggest that pre-college students can evaluate teachers in a reliable and consistent manner (Amatoro, 1954; Christensen, 1960). The validity of student reports is supported by a study which found the ratings of

eleventh and twelfth graders to be quite similar to those of experts (Bryan, 1966). Haak et al. (1972) report that ratings of older students are remarkably reliable. Although the reliability of elementary school student ratings has not been as thoroughly researched, Haak et al. (1972) summarized studies which indicate teacher ratings by younger students (down to grades 2-3) are valid; in addition they cite six studies which indicate that elementary student reports of peers are quite reliable.

In summary, research literature and professional experience suggest that student reports and evaluations of teachers, particularly in reference to discrete and visible behaviors, are potentially an important source of information for teacher evaluation. It is also evident that still more research is needed in this area in order to bring student reports into teacher evaluation practice.

PEER REVIEW

Teacher peer review brings the expertise and experience of the profession into evaluation as does no other assessment technique. Yet, it is one of the more undeveloped and under-researched areas of teacher evaluation (Batista, 1976). Teacher colleagues are familiar with school goals, priorities, values, and problems (Ryans, 1975) and are aware of the actual demands, limitations, and opportunities which face classroom teachers. They are in a position to address both the quality of teaching and the real limitations of actual teaching situations. The present difficulties with peer review in teacher evaluation are considerable. Chiefly, they stem from lack of reliable procedures, credibility to outside audiences, and teacher preparation for peer evaluation. Problems also arise because peer review is not an established and administrator-sanctioned

part of educational systems.

Arguments for the development and use of peer review are compelling. Teachers in the same subject area can give highly specific feedback. Colleague judgments about academic quality, currency of information, and scholarly organization provide additional perspective to student ratings and other evidence. Experience with how classes work and how children learn permits judgments which are realistic and pertinent. Peer review can be healthy for the professional life of teachers; it encourages professional behavior and helps lessen the professional isolation which occurs in teaching (Lortie, 1974).

The bulk of research on peer review has focused on one topic, the efficacy of classroom visits. Teacher visits are as unreliable as are those of administrators and other supervisors (see the section on Classroom Visits). Studies suggest that the unreliability is due to the few number of observations, judgments based on political considerations or friendships, and overreliance on style preferences which have little to do with the objectives of teaching (Scriven, 1981). As Centra (1977) has stated: "Colleague ratings of teaching effectiveness based primarily on classroom observation would in most instances not be reliable enough to use in making decisions on retention and promotion - at least not without faculty members investing much more time in visitations or in training sessions."

A number of writers contend that peer review is best done by considering materials which are used in the classroom. French-Lazovik (1981) described college level systems which call for syllabi, study guides, reading lists, assignments, texts, and course outlines to be used as evidence for peer judgments about: (a) quality of materials, (b) kinds of

intellectual tasks, and (c) how knowledgeable the instructor is about the topic. Scriven (1981) suggested that, in addition, tests and other feedback given to students (e.g., comments on papers and exams) be used as evidence of quality. He also suggested that fairness, quality of teacher assessments, and evidence of unusually bad practices be looked for. Review of instructional materials has the advantages of logistical practicality, lack of focus on classroom style, commonality of formats, and potential to examine discrete elements of teaching which are important indicators of quality. Because of the proximity of the evaluators to the teaching situation, peer review might also provide a workable opportunity to include student achievement data in teacher evaluation.

Some methodological problems with peer review can be solved with increased attention to the standardization of these procedures. Development of uniform procedures for materials review can provide an effective tool for teacher evaluation. Credibility of peer review can be established with the use of corroborating data (e.g., student reports and systematic observation). Teacher bias (Lewis, 1975; Batista, 1976; Stumpf, 1980) can be attenuated if the procedure is seen by the profession as a fair and supportive contribution. In order to implement peer review systems it will be necessary for administrators to review power relationships which currently exist in schools. While the efforts and expense of developing peer review into an accepted teacher evaluation technique are apparent, the payoff in improved practice and satisfaction makes them worthwhile.

COMPETENCY-BASED TEACHER EVALUATION

Competency-based teacher evaluation (CBTE) is an approach which relies

on assessing the performance of a teacher on a given set of basic teaching skill components. Emphasis is placed upon demonstration of a person's capacity in each category of a system of teaching abilities. The component abilities are combinations of skills and understandings which, if performed with competence, are expected to result in effective teaching. The following three competency areas are taken from the 14 specified in the Georgia State System (Georgia State Department of Education, 1980):

> Organizes instruction to take into account individual differences among learners.
>
> Reinforces and encourages learner involvement in instruction.
>
> Demonstrates enthusiasm for teaching and learning and the subject being taught.

The CBTE idea is based on the following arguments:

> -educators can agree on a number of powerful[1] principles of effective teaching (Coker et al., 1980).
>
> -a conceptual framework of teaching is important to communicate, analyze, diagnose and monitor performance (Howsam & Houston, 1972).
>
> -if competencies are not all of what a teacher does, at least, (a) they are precursors of complete teaching, (b) the bulk of what is presently known about effective teaching can be represented, and (c) persons who demonstrably lack competencies in test situations should remediate or not teach.

These and other arguments have been developed by proponents of competency-based education (e.g., Heath & Nelson, 1974).

Competency-based education is <u>not</u> universally accepted by educators (Benham, 1981). There are a number of serious logical, empirical, and practical drawbacks. As stated by Travers (1981):

> The concept of teaching as an assembly of competencies

[1]predictive, explanatory, generalizable

> lacks substance at present. It has not led to the development of any defensible and usable set of criteria of teacher effectiveness. The approach has appeal, particularly to those who know little about what has, and has not, been established about the nature of teaching. For the latter reason; it has had political attractiveness and has found some acceptance among some members of state legislatures, who have then brought pressure to bear on state departments of education to apply the concept to teacher certification, teacher evaluation, and teacher education. (p. 21).

Critique of CBTE

A critique of CBTE is based on five main arguments, which are discussed in this section:

1. Actual teaching performance is not merely the sum of distinct competencies.

2. Generic competencies are greatly limited by the context-dependency of actual teaching and learning.

3. While there is agreement on many specific relationships which exist between teacher performance and student learning, there is not agreement on a system or set of components which describe the entirety of teaching performance.

4. Competencies are not the same as the process-product research findings on which some persons have claimed CBTE is based; competency systems have not been empirically verified.

5. Not all of teaching can be reduced to a competency framework.

The act of teaching is one of implementing a plan in terms of an actual student population. In doing this, the teacher must adjust her intents and actions in relation to the group. The focus of the teacher is not on specific strategies but on a combination of them which best accommodates the plan and the actual teaching situation. As described by Brophy and Evertson (1976):

> Effective teaching is not simply a matter of implementing a small number of basic teaching skills. Instead, effective teaching requires the ability to implement a very large number of diagnostic,

> instructional, managerial, and therapeutic skills, tailoring behavior in specific contexts and situations to the specific needs of the moment. Effective teachers not only must be able to do a large number of things; they also must be able to recognize which of the many things they know how to do applies at a given moment and be able to follow through by performing the behavior effectively. (p. 139).

This interaction between plan and actual situation has an analogy in many team sports. When a particular game plan works, it is effective; when it does not, modifications must be made to fit the situation.

Competency-based evaluation implies that minimal performance of each of a collection of discrete capacities is adequate. In reality, many effective teaching practices (such as clarity and supportiveness) have a curvilinear rather than a linear effect on learning. Some demonstration of the ability, at the proper time, enhances learning while too much of the same competency retards learning (Soar, 1973). Thus, teachers who score high on competency assessments may be miserable teachers because they do not alter their behavior when it is called for.

A second major problem of CBTE is that the generic categories in competency systems are greatly limited by the many context influences on teaching outcomes. Educational contexts which have been shown to alter the way in which competency should be performed include:

age of student
prior achievement
type of educational goal
size of class
general school morale
grouping patterns
socioeconomic status

No set of generic competencies holds over the range of actual conditions found in teaching. Thus, competency systems are ineffective in discerning actual effective teaching performances. Coker, Medley, and Soar (1980)

focused on 25 competencies (e.g., "uses nonverbal communication skills," "gives clear and explicit directions," and "uses student feedback to modify teaching practices") which were systematically observed in 100 Georgia classrooms over a two-year period. Only six of the 25 competencies were found to be positively related to achievement gains, and five were positively related to student self-concept gains. Five were negatively related to achievement gains, and five were negatively related to self-concept gains. Several others were negatively related at some levels (grades) and positively related at others. These results strongly support the idea of the context dependency of learning; the validity of generic competencies across all grade levels was strongly questioned.

Actual experience gained with competency-based educational systems has not produced backing for their adoption in specific applications such as teacher evaluation. Heath and Nelson (1974) reported that research has not indicated that competency-based systems result in significant educational gains. Woditsch (1978) examined a number of competency-based systems and reported that the actual instruction and materials did not differ from more conventional programs and the same was true with the results. Only an increase in clarity of goals, relative to other educational approaches, was noted.

Often, advocates of CBTE refer to an empirical, or research, basis for competency systems. However, this research backing is indirect. Process-product studies, in which correlational relationships between teacher behaviors and learning outcomes are sought (e.g., Soar, 1973), are often cited. These studies report the effects of specific behaviors in given contexts and are not intended to be parts of generalizable competency systems.

In addition to their proven lack of validity, CBTE systems suffer from another major problem--cost. They are expensive to implement, as evidenced by the cost of the Georgia system, and their ability to eliminate incompetent teachers is unproven. It should be noted that the major emphasis of these systems to date has been elimination of incompetent teachers with little or no attention to improvement or recognition of superior teaching. This emphasis on the negative aspects of teacher evaluation coupled with validity and cost considerations makes this form of teacher evaluation less attractive than a number of alternatives.

SYSTEMATIC OBSERVATION

Evaluating teaching through systematic observation is a process whereby the actual classroom performance of a teacher is documented and analyzed in detail. While on the surface this appears to be a crucial and obvious source of information about the quality of teachers, in practice systematic observation is difficult and expensive to do well, and somewhat limited in scope.

What makes observation systematic?

Classroom observation is systematic when it fairly represents what goes on in the classroom, can be agreed upon by knowledgeable persons, and when the content of the observations are defensible in terms of their educational importance. Specifically, this means that the following five practices or limitations are in effect:

1. The observer is trained in the techniques of observation and is checked for actual reliability in practice (see Flanders, 1970).

2. The number and timing of visits are planned to insure a fair and reliable sample of classroom time and events (this may involve approximately eight sessions--depending on what is

observed and how variable the activities in the classroom are).

3. The focus of observation is limited to a specific number of visible categories which have proved to be reliably observable in practice. Since trained observers have limits of what they can pay attention to, their attention needs to be focused.

4. The recording system (checklists, entry forms, scoring) needs to be systematic, verifiable, and permanent.

5. Data should be analyzed with a single, coherent conceptual framework which has been systematically validated to show its links with important features (e.g., student learning, school needs, legal expectations).

Absence of any one or more of these features seriously threatens the fairness, accuracy, or importance of systematic observation systems.

What should be observed?

There is no single, simple set of practices or events by which we can judge the value of teacher performance. This is because teachers perform a number of different roles varying from nurturant, to instructional, to managerial. There are, however, a good number of specific criteria which have been consistently shown to be of value for different teaching situations. The context of the teaching situation can be analyzed, and a useable and satisfactory set of observations for that setting determined. Context differences which must be examined before selecting the observation categories include the following:

Type of learning goal (e.g., achievement or creativity)
Subject matter (e.g., art or mathematics)
Instructional task (e.g., seatwork or chemistry experiment)
Time of year (e.g., first month of class or end of school year)
Students (e.g., age, economic background, prior learning)
Number of students (e.g., 12 or 40)
Amount of student participation (e.g., individual practice or group discussion)
Resources available (e.g., media, hands-on materials, paper-pencil)

Given that the above context or situational variables are taken into account, there are a number of teacher performance variables which can be

observed reliably and validly. One example is direct instruction of academic material for achievement learning, such as occurs in the teaching of basic math skills. Other kinds of instruction, for example, the development of positive attitudes in a literature class or problem-solving in government, call for different observational strategies.

Much of what teachers are expected to do falls in the category of direct instruction of academic material for achievement learning. This includes, for example, much of the content of chemistry, reading for comprehension, and computational skills in mathematics. Research has shown that teachers produce greater learning in students when they (a) effectively use time well, (b) perform direct instruction, and (c) manage learning productively.

Effective use of time is a very important teacher variable which has been found to affect student learning (Rosenshine, 1979; Fisher, et al., 1978). Fisher labeled this variable Academic Learning Time (ALT) and investigated three aspects of it. The first is time allocated to academic learning. Simply put, student learning is increased if more time is actually spent on the subject matter rather than on organizing, ordering, general discussing, or decision making. Powell and Dishaw (1980) reported that allocated time in second grade classrooms that they observed varied from 62 to 123 minutes per day and for fifth graders from 71 to 134 minutes per day. Clearly, some teachers are more adept at providing the time necessary for learning which is essential to students. The second part of ALT is engaged time, that time in which students are actively involved in learning the material. Powell and Dishaw reported engaged times from 38 to 98 minutes per day for second grade classrooms and a range of 49 to 105 minutes per day for fifth grades. Again, there

are important and distinct differences in the amount of time teachers provide for students to actually be at the work of learning academic material. The final part of ALT which has been shown to influence student learning, is the amount of success (being correct or accurate) that students have. It has been demonstrated that students learn more academic material when they can practice it with success. Teachers have been found to be characteristically different in the successful practice they provide for students, which in turn influences student learning.

Another area of teacher performance which makes a difference in academic learning is that of active teaching or direct instruction (Good & Brophy, 1978). Direct instruction refers to a teacher's performance of the following in a smooth, consistent, and understandable manner:

-clear goals, understood by students
-actively focused on getting tasks done
-frequent monitoring of progress
-illustrations, examples of how to do the work
-opportunity for students to practice and recite
-difficulty level controlled for interest and success
-much non-judgmental feedback, evaluation, information.

A third promising area of systematic observation is managing activity during instruction. These behaviors include the following:

-clear focus on some goals
-task orientation to procedures
-students involved in learning
-pace brisk but not exhausting
-optimistic, expectant of success
-consistent management
-consistent treatment of high and low achievers.

These characteristics within a room have been demonstrated to support academic learning (Good & Brophy, 1978).

Use of systematic observation in teacher evaluation

Systematic observation provides a great deal of information about how well a teacher is working but has significant limitations for evaluating

overall teacher quality. First of all, what has been observed to date in teaching, namely academic type learning, is very important--but it is not the entirety of what a teacher is or does. Second, performance judgments are context-dependent; the type of learning, the nature of the students, and the other context variables described in the above section all need to be taken into account. Thus, comparison, which is a key feature of evaluation, is very difficult to set up. It is not often that teachers are in situations which are comparable. A third problem with systematic observation when it is used in an evaluation system is that it, like other techniques, can be disruptive of individual teaching patterns; teachers can be disrupted into attending to a system rather than paying attention to their own developed patterns. Finally, it should be recalled that a good systematic observation system is expensive and logistically complicated. This latter consideration needs to be looked at in cost-benefit terms.

Systematic observation is a powerful tool. Its use in formative evaluation is clear. Its potential for summative evaluation is not as clear; at least it would have to be combined with other kinds of data. If it is used for formative purposes, a support system is also needed. That is, the information should be given to the teacher, and then in-service follow-up provided to help the teacher alter practice and acquire skills to improve performance.

SECTION III

LEGAL ISSUES IN TEACHER EVALUATION

TEACHER EVALUATION - LEGAL ISSUES

A discussion of teacher evaluation practices would be incomplete without some consideration of the legal issues involved, since teacher evaluation systems inevitably produce many instances where questions of fairness and judgment exist. The trend in education is clear; clients as well as educators within the profession are turning to the courts for the settlement of educational controversies (Frances & Stacy, 1977; Joyce, 1978). The area of teacher evaluation will not be an exception.

A major issue in the implementation of any teacher evaluation system is due process, protected by the 14th Amendment to the Constitution. This amendment guarantees procedural due process, which includes the right of notice of dismissal, a hearing, and in some instances a statement of the reasons for dismissal (Centra, 1980). There has been a tendency for the courts to strictly apply the procedural requirements of teacher evaluation laws. In addition when district level policies exist, they must be followed closely and administered in a non-biased fashion (Zirkel, 1979-80). In non-educational cases, courts have rendered decisions which do not support the inappropriate use of performance evaluations in instances where: 1) ratings were based on subjective or vague factors; 2) observational ratings did not indicate an adequate sampling of behavior, or there was evidence to indicate rater bias; and 3) standard conditions were not employed for the collection and scoring of ratings (Griggs et al. v. Duke Power Co., 1970).

The courts have been fairly rigorous in the interpretation of the concept of due process to educational cases. Dismissal charges against teachers in the state of Pennsylvania were <u>not</u> sustained in cases where the rating systems were not strictly followed, where the evaluation form

did not contain unsatisfactory ratings and where the required anecdotal records were not provided. In addition, courts in various jurisdictions have overturned dismissal decisions based on unsatisfactory evaluations due to failure to provide written warnings about remediable teaching deficiencies (Zirkel, 1979-80).

The difficulty in implementing teacher evaluation systems for the purpose of teacher dismissal can be seen in a case study of Pennsylvania cases. Zirkel reports

> ....in Pennsylvania which has probably the most lengthy and well-developed legal history concerning teacher evaluation, only about 100 teachers have been charged with incompetence by local boards since 1940, averaging 2.7 per year, and the charges have been upheld against only slightly above 50 percent of the teachers (1979-80, p. 21).

One response to the difficulties involved in implementing the results of a teacher evaluation system is to develop more detailed and specific evaluation procedures. But even this action can be counter-productive; the more detailed the procedures, the greater the possibility that some procedural shortcoming will occur.

> Paradoxically, if an institution's personnel practices are vague or unspecified, it is more difficult for faculty members to challenge decisions on specific procedural grounds (Cohen, 1981, p. 39).

Another major legal issue in the implementation of teacher evaluation systems involves the validity of the systems themselves. Here the courts have been much more willing to defer to the discretion of school authorities. For example, despite the problems inherent in the use of standardized test scores for teacher evaluation, the courts have not been willing to overturn dismissal cases based upon these types of data. In reaching this conclusion Zirkel (1979-80) cautioned that the specifics of a case

such as whether the teachers were tenured or not and the specific state statutes involved could influence future courts' decisions.

Within broad general limits the courts do not appear to be interested in determining the particular methods of evaluation or the criteria that are applied. The courts do, however, expect the evidence obtained to be valid, i.e. job related and non-discriminatory. In this regard Cohen (1980) offers both general and specific legal advice to those who develop teacher evaluation systems. In developing these systems, administrators should consider the evidence that needs to be advanced in a court case to defend the validity of assessment methods used. Specifically he cautions "...that rating scales or evaluation systems that include such criteria as the teacher's appearance, neatness or sense of humor are questionable in any case" (1980, p. 145).

SECTION IV

DIRECTIONS FOR DEVELOPMENT OF TEACHER EVALUATION

As described in the Introduction to this report, the present appears to be a good time to enhance efforts for research and development in teacher evaluation. The purposes, needs, and techniques for evaluation have become clearer. A variety of approaches and strategies may be combined to result in evaluation which satisfies a good number of audiences.

Research and development in teacher evaluation will require cooperative efforts of universities, school districts, state school boards, and public interest groups. A realistic time-frame and set of expectations must be established: progress in teacher evaluation will take time, trial and error, collaboration, and some additional money. The efforts and expenditures can be expected to be well worthwhile to the various audiences of teacher evaluation, whose present dissatisfactions are obvious.

It is necessary to involve teachers at the outset and throughout any teacher evaluation study project. First, it would be difficult to increase teacher satisfaction without knowing more about their roles and needs. Second, benefits of evaluation data for other audiences (e.g., lay public and universities) need to be coordinated with teacher benefits. Finally, successful development requires the teacher cooperation that "top-down" educational projects rarely receive from participants.

The implementation of any teacher evaluation system must also consider other factors. These include an analysis of the cost-benefits involved, the state of the art in different areas of teacher evaluation, and the kind of data provided. With these ideas in mind, the authors recommend the following as areas of potentially promising practices.

A. Peer Review. The active involvement of teachers in the evaluation process in addition to providing valuable

evaluation information, would result in increased professionalism and responsibility for practicing teachers.

B. Student reports. Reliable and valid information from students can be systematically used in a fair and informative manner, and can provide a unique facet in a total teacher evaluation system.

C. Systematic observation. If done correctly, systematic observation can provide valuable formative and summative information.

D. Academic screening. A more rigorous screening process through the use of standardized aptitude and achievement measures would help to insure the quality of teachers entering the profession.

E. Evaluation systems for first and second year teachers which are cooperatively managed by districts, universities, teacher organizations, and state departments of education.

REFERENCES

Aleamoni, L. Typical faculty concerns about student evaluation of instruction. National Association of Colleges and Teachers of Agriculture Journal, 1976, 20, 16-21.

Aleamoni, L. Development and factorial validation of the Arizona Course/Instructor Evaluation Questionnaire. Educational and Psychological Measurement, 1978, 38, 1063-1067.

Aleamoni, L. Student ratings of instruction. In J. Millman, (Ed.), Handbook of Teacher Evaluation. Beverly Hills, California: Sage Publications, 1981.

Aleamoni, L. and Hexner, P. A review of the research on student evaluation and a report on the effect of different sets of instruction on student course and instructor evaluation. Instructional Science, 1980, 9, 67-84.

Aleamoni, L. and Spencer, R. The Illinois course evaluation questionnaire: A description of its development and a report of some of its results. Educational and Psychological Measurement, 1973, 33, 669-684.

Amatora, M. Teacher ratings by younger pupils. Journal of Teacher Education, 1954, 5, 149-152.

Andrews, J., Blackman, C. and Mackey, J. Preservice performance and the national teacher exam. Phi Delta Kappan, 1980, 61, 358-359.

Batista, E. The place of colleague evaluation in the appraisal of college teaching. Research in Higher Education, 1976, 4, 257-271.

Benham, B. CBTE: Another educational edifice built on quicksand. The Teacher Educator, 1981, 17(1), 26-29.

Blackburn, R. and Clark, M. An assessment of faculty performance: Some correlates between administrators, colleagues, students and self ratings. Sociology of Education, 1975, 48, 242-256.

Borich, G. The Appraisal of Teaching: Concepts and Processes. Reading, Massachusetts: Addison-Wesley, 1977.

Brophy, J. and Evertson, C. Learning from Teaching: A Developmental Perspective. Boston: Allyn & Bacon, 1976.

Brophy, J. and Evertson, C. Context variables in teaching. Educational Psychologist, 1978, 12, 310-316.

Bryan, R. Teacher's image is stubbornly stable. Clearing House, 1966, 40, 459-461.

Carroll, J. Faculty self evaluation. In J. Millman, (Ed.), Handbook of Teacher Evaluation. Beverly Hills, California: Sage Publications, 1981.

Carter, W. An interpretive analysis of the teacher selection and evaluation process. (Pub. No. RE 97-804-61-05). Dallas: Dallas Independent School District, 1979.

Centra, J. Strategies for Improving College Teaching. Washington, D.C.: American Association for Higher Education, 1972.

Centra, J. Self-ratings of college teachers: A comparison with student ratings. Journal of Educational Measurement, 1973, 10, 287-295.

Centra, J. Colleagues as raters of classroom instruction. Journal of Higher Education, 1975, 46, 327-337.

Centra, J. Student ratings of instruction and their relationship to student learning. Research Bulletin 76-6. Princeton, New Jersey: Educational Testing Service, 1976.

Centra, J. The how and why of evaluating teaching. New Directions for Higher Education, 1977, 17, 93-106.

Centra, J. Determining faculty effectiveness. In J. Centra, (Ed.), Determining Faculty Effectiveness, San Francisco: Jossey-Bass, 1980.

Centra, J. and Creech, F. The relationship between student, teacher and course characteristics and student ratings of teacher effectiveness. P.R. 76-1, Princeton, N. J.: Educational Testing Service, 1976.

Christensen, C. Relationships between pupil achievement, pupil affect need, teacher warmth and teacher permissiveness. Journal of Educational Psychology, 1960, 51, 169-173.

Clark, M. and Blackburn, R. Assessment of faculty performance: Some correlates between self, colleagues, students and administrators. Ann Arbor: University of Michigan, Center for the Study of Higher Education, 1971.

Cohen, P. Student ratings of instruction and student achievement: A meta analysis of multisection validity studies. Review of Educational Research, 1981, 51, 281-310.

Coker, H., Medley, D. and Soar, R. How valid are expert opinions about effective teaching? Phi Delta Kappan, October, 1980, 131-149.

Cook, M. and Richards, H. Dimensions of principal and supervisor ratings of teacher behavior. Journal of Experimental Education, 1972, 41, 11-14.

Cornett, J. Effectiveness of three selective admissions criteria in predicting performance of first-year teachers. Journal of Educational Research, 1969, 62, 247-250.

Costin, F., Greenough, W., and Menges, R. Student ratings of college teaching: Reliability, validity and usefulness. Review of Educational Research, 1971, 41, 511-535.

Council for Basic Education. Testing teachers. Basic Education, 1978, 24, 3-6.

Educational Testing Service. Comparative Data Guide for the Student Instructional Report (1975-1976). College and University Programs, Princeton, New Jersey, 1975.

Evertson, C. and Holley, F. Classroom observation. In J. Millman, (Ed.), Handbook of Teacher Evaluation. Beverly Hills, California: Sage Publications, 1981.

Festinger, L. A. A theory of social comparison process. Human Relations, 1954, 7, 117-140.

Fisher, C., Filby, N., Marliave, R., Cahen, L., Dishaw, M., Moore, J. and Berliner, D. Teaching behaviors, academic learning time, and student achievement: Final report of Phase III-B, Beginning Teacher Evaluation Study. San Francisco: Far West Laboratory for Educational Research and Development, 1978.

Frances, S. O. and Stacey, C. Law and the sensual teacher. Phi Delta Kappan, 1977, 59, 98-102.

French-Lazovik, G. Documentary evidence in the evaluation of teaching. In J. Millman, (Ed.), Handbook of Teacher Evaluation. Beverly Hills, California: Sage Publications, 1981.

Gallup, G. The eleventh Gallup Poll of the public's attitudes toward the public schools. Phi Delta Kappan, 1979, 61, 33-45.

Georgia State Department of Education. Teacher Competencies for the Georgia State Evaluation System. Athens, Georgia: University of Georgia, 1980.

Glass, G. A review of three methods of determining teacher effectiveness. In H. Walberg (Ed.), Evaluating Educational Performance. Berkeley: McCutchen, 1974.

Good, T. and Brophy, J. Looking in Classrooms (2nd ed.). New York: Harper & Row, 1978.

Griggs, et al., v. Duke Power Co., 401 US 424, 1970.

Guthrie, E. The Evaluation of Teaching: A Progress Report. Seattle: University of Washington, 1954.

Guthrie, J. Survey of school effectiveness studies. In A. Mood (Ed.), Do Teachers Make a Difference? Washington, D.C.: U. S. Government Printing Office, 1970.

Haak, R., Kleiber, D. and Peck, R. Student Evaluation of Teacher Instrument II. Austin, Texas: R & D Center for Teacher Education, 1972.

Harris, W. Teacher command of subject matter. In J. Millman, (Ed.), Handbook of Teacher Evaluation. Beverly Hills, California: Sage Publications, 1981.

Heath, R. and Nelson, M. The research basis for performance-based teacher education. Review of Educational Research, 1974, 44, 463-484.

Hogan, T. Similarity of student ratings across instructors, courses and times. Research in Higher Education, 1973, 1, 149-154.

Howsam, R. and Houston, W. Competency Based Teacher Education. Palo Alto, California: Science Research Associates, 1972.

Ingils, C. Let's do away with teacher evaluation. The Clearing House, 1970, 44, 451-456.

Joyce, M. Law and the laboratory. The Science Teacher, 1978, 45, 23-25.

Kaplin, W. The Law of Higher Education: Legal Implications of Administrative Decision Making. San Francisco: Jossey-Bass, 1978.

Kauchak, D. and Eggen, P. A comparison of peer, self and administrator evaluations in university faculty members. Presented at Association for Supervision and Curriculum Development, Miami, 1976.

Kerlinger, F. Student evaluations of university professors. School and Society, 1971, 99, 353-356.

Lewis, L. Scaling the Ivory Tower: Merit and its Limits in Academic Careers. Baltimore: John Hopkins Press, 1975.

Lortie, D. Schoolteacher. Chicago: University of Chicago Press, 1974.

Maslaw, A. and Zimmerman, W. College teaching ability, scholarly activity, and personality. Journal of Educational Psychology, 1956, 47, 185-189.

McNeil, J. and Popham, W. The assessment of teacher competence. In R. M. Travers (Ed.), Second Handbook of Research on Teaching. Chicago: Rand McNally, 1973, 131-147.

Medley, D. and Mitzel, H. Some behavioral correlates of teacher effectiveness. Journal of Educational Psychology, 1959, 50, 239-246.

Millman, J., Ed. Handbook of Teacher Evaluation. Beverly Hills, California: Sage Publications, 1981.

Mitchell, R. Testing the teachers: The Dallas experiment. School Leader, 1979, 8, 20-23.

Murray, H. The validity of student ratings of teaching ability. Paper presented at the Canadian Psychological Association, Montreal, 1972.

McKeachie, W. Student ratings of faculty: A response. Academe, 1979 65, 384-397.

Northen, E. The trend toward competency testing of teachers. Phi Delta Kappan, 1980, 61, 359.

Pambookian, H. Initial level of student evaluation of instruction as a source of influence on instructor change after feedback. Journal of Educational Psychology, 1974, 66, 52-56.

Pambookian, H. Discrepancy between instructor and student evaluations of instruction: Effect on instructor. Instructional Science, 1975, 5, 63-75.

Patton, R. and Desena, P. Identification through student opinion of motivating and nonmotivating qualities of teachers. Journal of Teacher Education, 1966, 17, 41-45.

Perry, R., Abrami, P. and Leventhal, L. Educational seduction: The effect of instructor expressiveness and lecture content on student ratings and achievement. Journal of Educational Psychology, 1979, 71, 107-116.

Peterson, K. and Yaakobi, D. Israeli science students and teacher perceptions of classroom role performance: Concepts, reports, and adequacy. Science Education, 1980, 64(5), 661-669.

Popham, W. Performance tests of teaching proficiency: Rationale, development, and validation. American Educational Research Journal, 1971, 8, 105-117.

Powell, M. and Dishaw, M. A realistic picture of reading instructional time. Reading Research Quarterly, 1980, 16.

Rencher, A., Wadham, R. and Young, J. A discriminant analysis of four levels of teacher competence. Journal of Experimental Education, 1978, 46(3), 46-51.

Rosenbloom, P. Characteristics of mathematics teachers that affect students' learning. ERIC Document E D021707, 1966.

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, 40(5), 647-662.

Rosenshine, B. Academic engaged time, content covered, and direct instruction. Journal of Education, 1978, 160(3), 38-66.

Schalock, D. From research to practice: The dilemma for teacher education. Address at Oregon Education Research Association, Otter Crest, Oregon, October 30-31, 1981.

Schmid, J. Factor analysis of the teaching complex. Wisconsin Studies of the Measurement and Prediction of Teacher Effectiveness. Madison, Wisconsin: Dembar Publications, 1968.

Scriven, M. The evaluation of educational goals, instructional procedures, and outcomes. ERIC Document ED 079 394, 1973.

Scriven, M. The evaluation of teachers and teaching. California Journal of Educational Research, 1974, 24(3), 109-118.

Scriven, M. The evaluation of college teaching. National Council of States on Inservice Education, June, 1980, 9-15.

Scriven, M. Summative teacher evaluation. In J. Millman, (Ed.), Handbook of Teacher Evaluation. Beverly Hills, California: Sage Publications, 1981.

Simun, P. and Asher, J. The relationship of variables in undergraduate school and school administrator's ratings of first year teachers. Journal of Teacher Education, 1964, 16, 293-302.

Soar, R. Teacher assessment problems and possibilities. Journal of Teacher Education, 1973, 24, 205-212.

Stumpf, W. Peer review. Science, 1980, 207, 822.

Sullivan, A. and Skanes, G. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 584-590.

Travers, R. Criteria of good teaching. In J. Millman, (Ed.), Handbook of Teacher Evaluation. Beverly Hills, California: Sage Publications, 1981.

Tuckman, B. and Oliver, W. Effectiveness of feedback to teachers as a function of source. Journal of Educational Psychology, 1968, 59(4), 297-301.

Weaver, W. In search of quality: The need for talent teaching. Phi Delta Kappan, 1979, 61, 29-46.

Weiner, B. and Kukla, A. An attributional analysis of achievement motivation. Journal of Personality and Social Psychology, 1970, 15, 1-20.

Woditsch, G. Specifying and achieving competencies. In O. Milton, (Ed.), On College Teaching. San Francisco: Jossey-Bass, 1978.

Zirkel, P. Teacher evaluation: A legal overview. Action in Teacher Education, 1979-80, 2, 17-25.